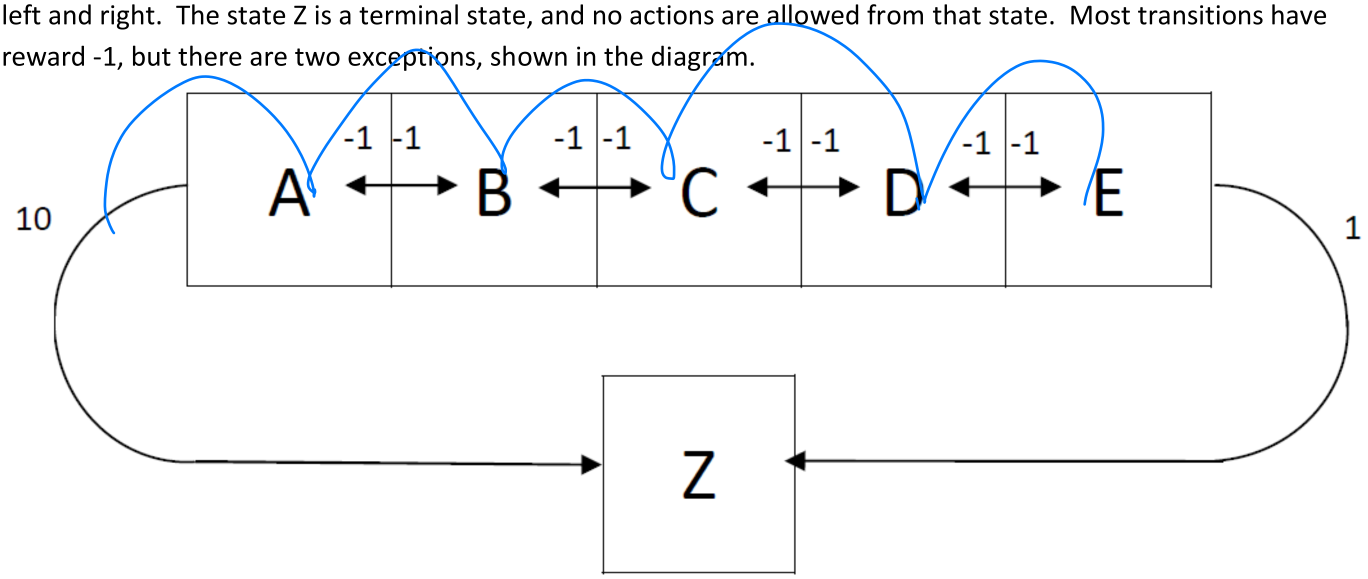
Intro to AI; Univ. of Washington. Your Name: Travis Xie

Markov Decision Process Policies, Values and Q-Values

Consider the MDP shown in the state-transition diagram below. There are six states and two actions {L, R} meaning left and right. The state Z is a terminal state, and no actions are allowed from that state. Most transitions have reward -1, but there are two exceptions, shown in the diagram.



1. One possible policy is π_1 = {(A,L), (B, L), (C, L), (D, L), (E, L)}, which can be written more concisely as π_1 = LLLLL. What are two other possible policies? π_2 = RRRRR , π_3 = LLLLR

2. How many possible policies are there for this MDP? $2^5 = 32$

Intro to AI; Univ. of Washington. Your Name: Travis Xie

3. Assume that T(s, *a*, s')=1 for each edge (s, s') in direction *a* shown in the diagram and T(s, *a*, s')=0 for all other triples. Assume γ = 1. Then write in the Q values shown. $Q_k(s_0, a)$ is the total reward expected given that the agent starts in state s_0 and has committed to taking action *a* and proceeds for k transitions, with all transitions after the first one using the optimal action (considering the number of transitions left) at each state reached.

	A	B	C	D	E	Z
$Q_0(s, L)$	0.0	0.0	0.0	0.0	0.0	-
$Q_0(s, R)$	0.0	0.0	0.0	0.0	0.0	-
$Q_1(s, L)$	10	-1	-1	-1	-1	-
$Q_1(s, R)$	-1	-1	-1	-1	1	-
$Q_2(s, L)$	10	9	-2	-2	-2	-
$Q_2(s, R)$	-2	-2	-2	0	1	-
$Q_3(s, L)$	10	9	8	-3	-1	-
$Q_3(s, R)$	8	-3	-1	0	1	-
$Q_4(s, L)$	10	9	8	7	-1	-
$Q_4(s, R)$	8	7	-1	0	1	-
$Q_5(s, L)$	10	9	8	7	6	-
$Q_5(s, R)$	8	7	6	0	1	-
$Q^*(s, L)$	10	9	8	7	6	-
$Q^*(s, R)$	8	7	6	5	1	-

4. Show the values $V_k(s)$. This is the expected total reward starting at state s and taking k transitions, each according to the optimal action at each state reached (including at state s).

	A	B	C	D	E	Z
$V_0(s)$	0.0	0.0	0.0	0.0	0.0	-
$V_1(s)$	10	-1	-1	-1	1	-
$V_2(s)$	10	9	-2	0	1	-
$V_3(s)$	10	9	8	0	1	-
$V_4(s)$	10	9	8	7	1	-
$V_5(s)$	10	9	8	7	6	-
$V^*(s)$	10	9	8	7	6	-

5. What would be the values $V_1(s)$ and $V_2(s)$ if T(s, a, s')=0.5 on each edge, meaning actions L and R amount to the same thing – a coin flip on whether the action actually takes you west or east?

	A	B	C	D	E	Z
$V_0(s)$	0.0	0.0	0.0	0.0	0.0	-
$V_1(s)$	4.5	-1	-1	-1	0	-
$V_2(s)$	4	-2	-2	-1.5	-0.5	-

6. What if instead of no discounting, we have γ = 0.5 (with deterministic actions again)? What would Q*(E, L) be? This is asking what the expected total discounted reward would be in you start in state E, commit to going left on your first action, and then act optimally. The discounting means that on the ith transition, your reward is multiplied by γ^{i-1}.

$Q^*(E, L) = -1.25$

1: $\overset{L}{(-1)} + \overset{R}{(-1)} \times 0.5 + \overset{L}{1} \times 0.5^2 = -1.25$

2: $\underset{L}{(-1)} + \underset{L}{(-1) \times 0.5} + \underset{L}{(-1) \times 0.5^2} + \underset{L}{(-1) \times 0.5^3} + \underset{L}{10 \times (0.5)^4} = -1.25$